Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung:
Verteilungen stetiger Zufallsvariablen

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

„. . . if the number of observations is large, the distribution of θ in repeated sampling tends to be, and for practical purposes is actually normal.“

Neyman, Jerzy: On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, in: Journal of the Royal Statistical Society, Series A 97, 1934, S. 558-625, hier S. 566.

Stetige Zufallsvariablen

Stetige Zufallsvariablen können im Prinzip unendlich viele Werte (u. U. innerhalb eines Bereichs annehmen). Daher können sie nur mit Hilfe der Infinitesimalrechnung traktiert werden.

Allerdings kommt man um stetige Zufallsvariablen nicht herum, da manche Stichprobenergebnisse (z. B. Mittelwerte) eben stetig sind. Zudem können eigentlich diskrete Verteilungen gut durch stetige Verteilungen approximiert werden, jedenfalls bei großen Stichproben. Gleichzeitig ist die Anwendung diskreter Verteilungen zwar im Prinzip einfach, aber mühsam (Ausrechnen vieler Einzelwahrscheinlichkeiten).

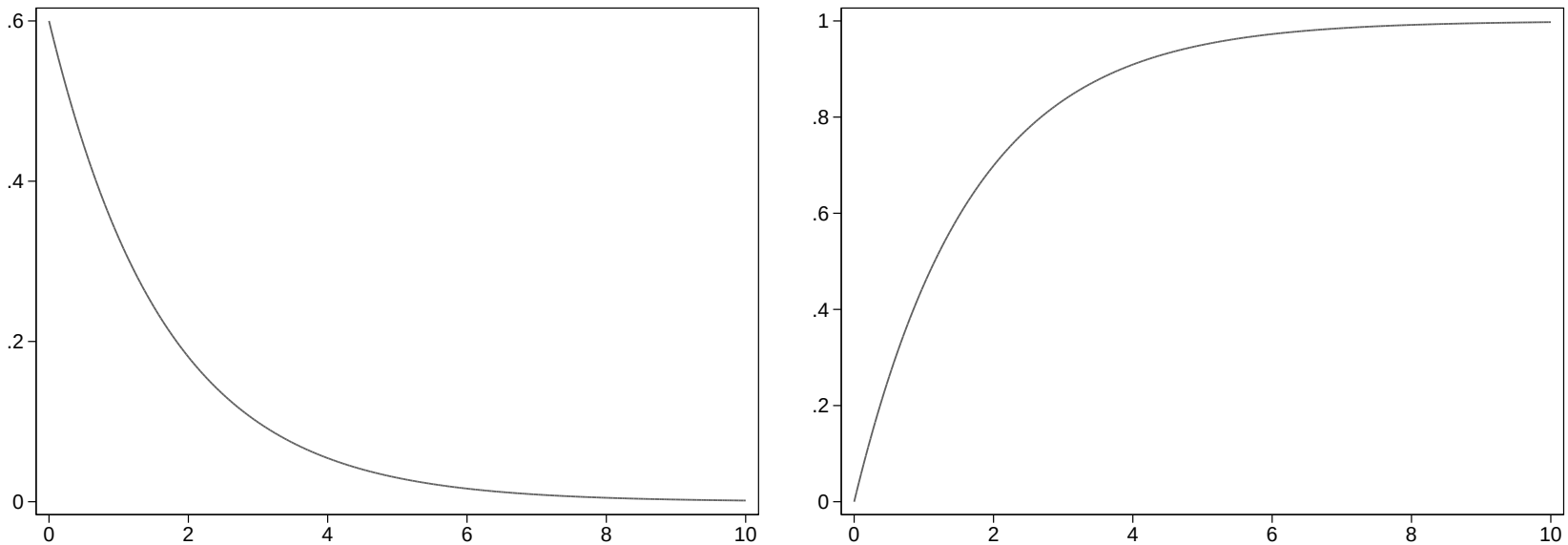
Und da stetige Verteilungen

- in Tabellen verfügbar bzw.
- in Statistik-Software implementiert sind

ist für die praktische Anwendung keine Infinitesimalrechnung nötig.

Beispiel: Exponentialverteilung

Eine Variable mit $f(x) = \lambda e^{-\lambda x}$ und $F(x) = 1 - e^{-\lambda x}$ für $x \geq 0$ heißt exponentialverteilt mit Parameter λ.



Dichtefunktion (links) und Verteilungsfunktion (rechts) einer Exponentialverteilung mit $\lambda = 0{,}6$.

Beschreibung stetiger Verteilungen

Bei stetigen Verteilungen ist die Wahrscheinlichkeit eines einzelnen Wertes nicht definiert (genauer gesagt: sie beträgt Null). Es können nur Wahrscheinlichkeiten in einem Intervall über die Bestimmung des Integrals über dieses Intervall berechnet werden.

Statt der Wahrscheinlichkeitsfunktion gibt es eine Dichtefunktion:

$$P(a \leqslant X \leqslant b) = \int_a^b f(x)dx$$

Die Verteilungsfunktion lautet allgemein:

$$F(x) = P(-\infty \leqslant X \leqslant x) = \int_{-\infty}^{x} f(u)du$$

Stetige Zufallsvariablen: E(X) und Var(X)

Erwartungswert und Varianz lassen sich ebenfalls nur unter Zuhilfenahme der Infinitesimalrechnung, aber in Analogie zu diskreten Variablen beschreiben:

Erwartungswert:

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Varianz:

$$\sigma_x^2 = Var(X) = \int_{-\infty}^{\infty} (x - \mu_x)^2 \cdot f(x) dx$$

Die Normalverteilung: Allgemein

Die Normalverteilung ist die wichtigste Verteilung der Statistik. Gründe hierfür:

- Empirische Verteilung: Viele Merkmale sind normalverteilt (Körpergröße, Gewicht) oder werden so konstruiert, dass sie normalverteilt sind (Tests).
- Fehlerverteilung: Zufällige Messfehler folgen einer Normalverteilung.
- Grundlegende Verteilung für die Inferenzstatistik: Bei großen Stichproben nähern sich andere Verteilungen der Normalverteilung an; gleichzeitig ist die Normalverteilung „Mutter“ einiger anderer wichtiger Verteilungen.

Die Normalverteilung

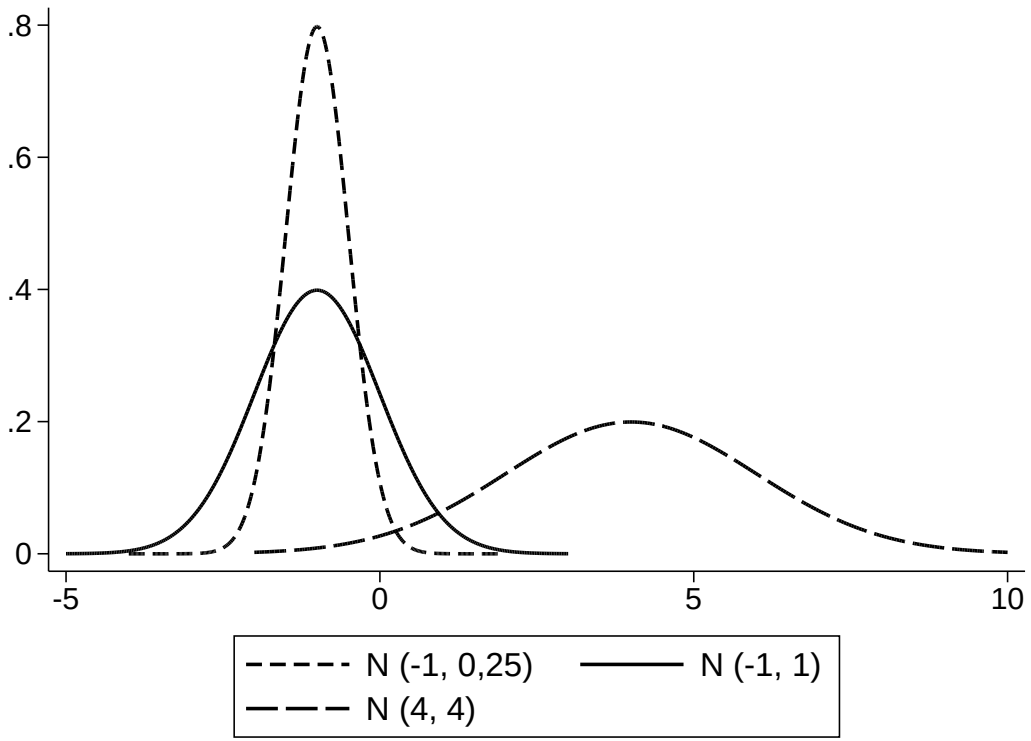
Die Dichteverteilung der Normalverteilung:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Dabei ist μ der Erwartungswert der Verteilung, σ^2 ihre Varianz und π die Kreiszahl 3,14...

Die Normalverteilung visualisiert

Es gibt also viele Normalverteilungen mit unterschiedlichen Mittelwerten μ und Varianzen σ^2. Für alle gilt: Sie sind symmetrisch um μ, unimodal, mehr oder weniger glockenförmig. Die Wahrscheinlichkeitsdichte strebt asymptotisch gegen 0, wenn x gegen $-\infty$ bzw. $+\infty$ strebt.



Die Standardnormalverteilung

Durch Standardisierung wird die Normalverteilung in die Standardnormalverteilung N (0;1) überführt.

Unter Standardisierung (auch z-Transformation) versteht man folgende Transformation:

$$Z = \frac{X - \mu}{\sigma}$$

Es wird von jedem einzelnen Wert von X der Mittelwert von X (hier geschrieben als μ) abgezogen und das Resultat durch σ dividiert. Jede standardisierte Variable hat einen Mittelwert von 0 und eine Standardabweichung von 1.

Die Dichtefunktion der Normalverteilung vereinfacht sich so zu dem (nicht zu lernenden) Ausdruck

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}}$$

Normalverteilung und Standardnormalverteilung

Für die Normalverteilung gilt:

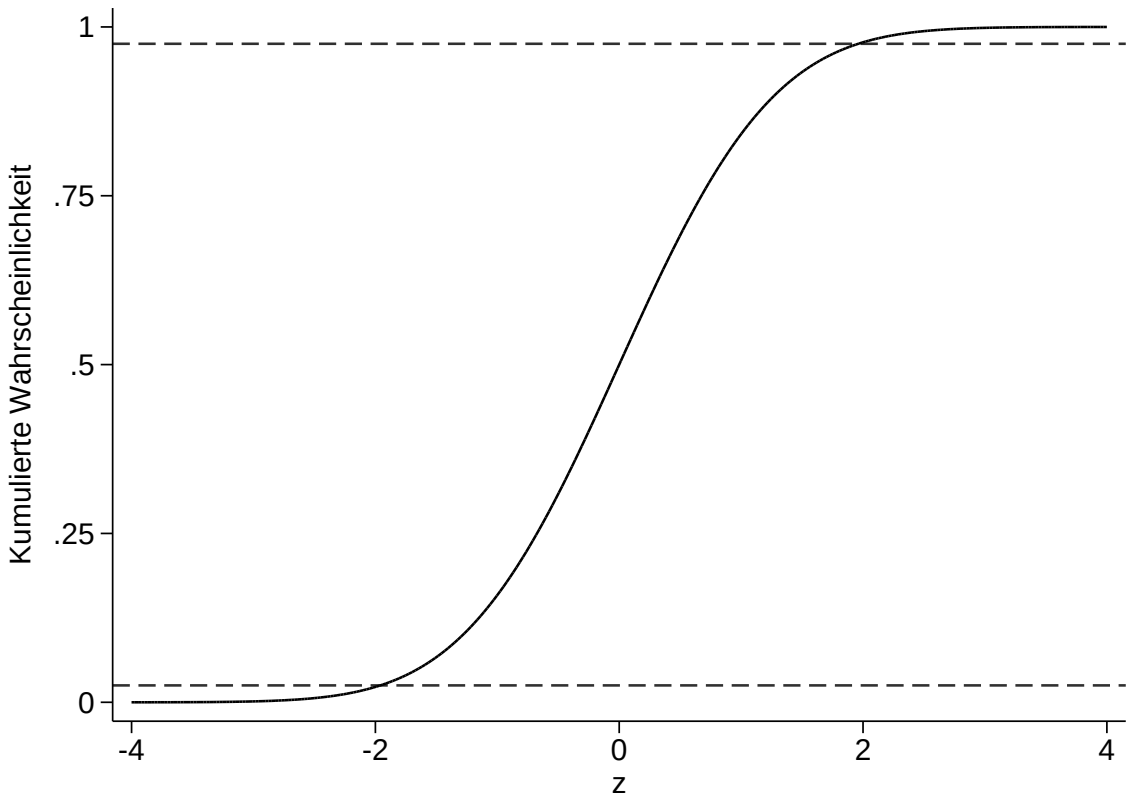
- Ca. 68 % der Werte liegen in einem Bereich von $\pm 1\,\sigma$ um den Mittelwert.
- Gut 95 % der Werte liegen in einem Bereich von $\pm 2\,\sigma$ um den Mittelwert.
- Ca. 99,7 % der Werte liegen in einem Bereich von $\pm 3\,\sigma$ um den Mittelwert.

Für die Standardnormalverteilung gilt entsprechend:

- Ca. 68 % der Werte liegen in einem Bereich von ± 1 um den Mittelwert.
- Gut 95 % der Werte liegen in einem Bereich von ± 2 um den Mittelwert.
- Ca. 99,7 % der Werte liegen in einem Bereich von ± 3 um den Mittelwert.

Standardnormalverteilung

Die Quantile der Standardnormalverteilung lassen sich grob aus der Verteilungsfunktion ablesen (gestrichelte Linien: 0,025-Quantil [$x \approx -2$] und 0,975-Quantil [$x \approx 2$]):



Einige Quantile der Standardnormalverteilung

Im Detail lassen sich die Quantile Tabellen entnehmen.

Die folgende Tabelle zeigt beispielsweise, dass der Wert des 0,01-Quantils (erstes Perzentil) $-2{,}325$ beträgt. Ein Prozent der Werte einer standardnormalverteilten Variablen ist also kleiner oder gleich $-2{,}325$, 99 Prozent sind größer oder gleich $-2{,}325$.

z	Quantil	z	Quantil
-3,000	0,0013	1,000	0,841
-2,325	0,01	1,282	0,90
-1,96	0,025	1,645	0,95
-1,645	0,05	1,96	0,975
-1,282	0,10	2,325	0,99
-1,000	0,159	3,000	0,9986
0	0,5		

Der zentrale Grenzwertsatz

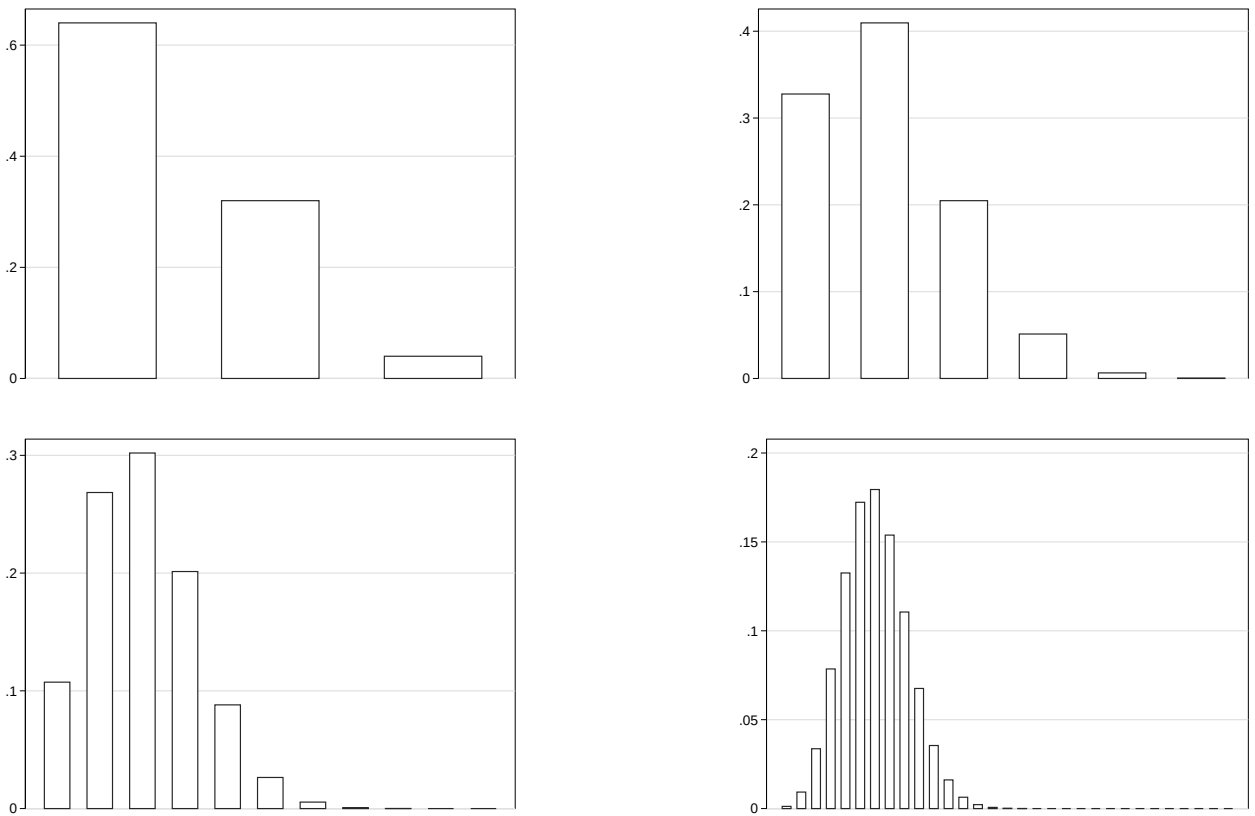
Die Verteilung der standardisierten Summe von n unabhängigen Zufallsvariablen, die alle die identische Wahrscheinlichkeitsverteilung haben, nähert sich mit steigender Stichprobengröße der Standardnormalverteilung an.

Daraus folgt u. a., dass Mittelwerte und Anteilswerte aus Zufallsstichproben bei „hinreichend großem“ n einer Normalverteilung folgen – auch wenn das Merkmal selbst in der Grundgesamtheit nicht normalverteilt ist.

„Hinreichend groß“ variiert je nach Umständen (Art des zugrunde liegenden Merkmals), mindestens gilt $n \geq 30$.

Der zentrale Grenzwertsatz illustriert

Eine diskrete und recht schief verteilte Variable (Binomialverteilung mit $\pi = 0{,}2$) wird mit zunehmendem n einer Normalverteilung ähnlicher (n = 2, 5, 10 und 30):



Die χ^2-Verteilung

Die Verteilung einer Summe unabhängiger quadrierter standardnormalverteilter Zufallsvariablen Z heißt χ^2-Verteilung:

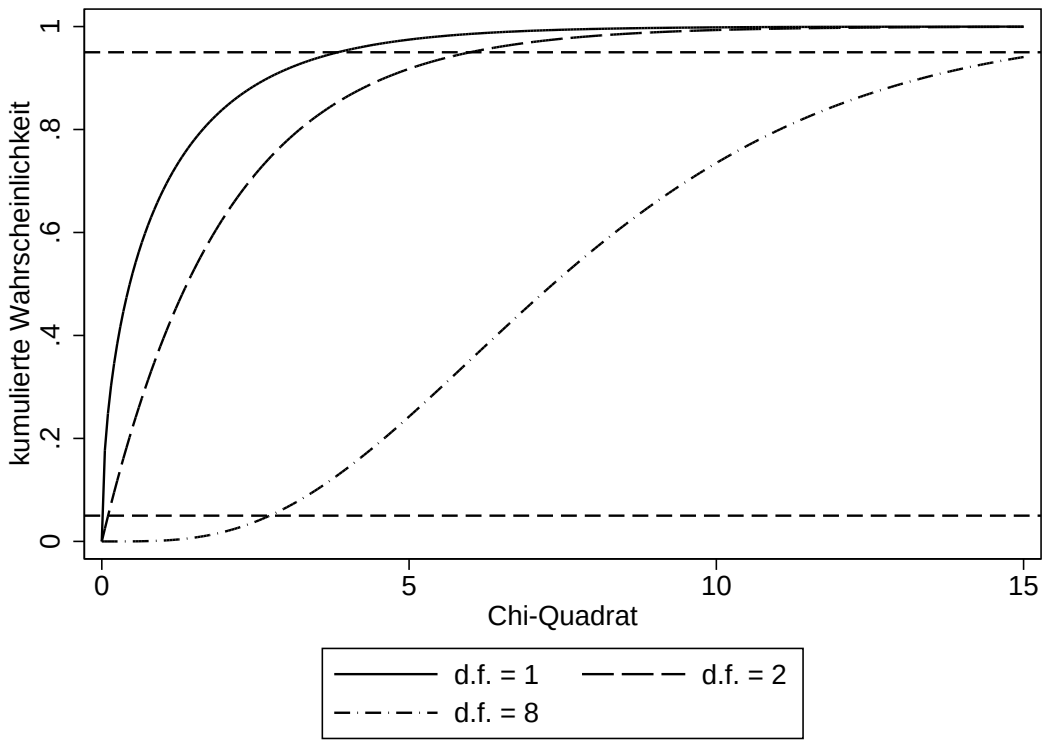
$$\chi^2_{df=n} = Z_1^2 + Z_2^2 + ... + Z_n^2$$

mit df=Zahl der Freiheitsgrade (degrees of freedom), d. h. Zahl der (von einander unabhängigen) Variablen Z.

Wichtige Anwendungsfälle: Tests in Kreuztabellen auf Überzufälligkeit; Likelihood-Ratio-Test in der Maximum-Likelihood-Schätzung.

Die χ^2-Verteilung illustriert

Verteilungsfunktionen einiger χ^2-Verteilungen (gestrichelt: 0,05- und 0,95-Quantil)



Die t-Verteilung

Die von „Student“ (Pseudonym für William S. Gosset) entwickelte t-Verteilung kann anstelle SNV herangezogen werden, wenn ein Merkmal in der Grundgesamtheit normalverteilt und die Varianz der Grundgesamtheit unbekannt ist.

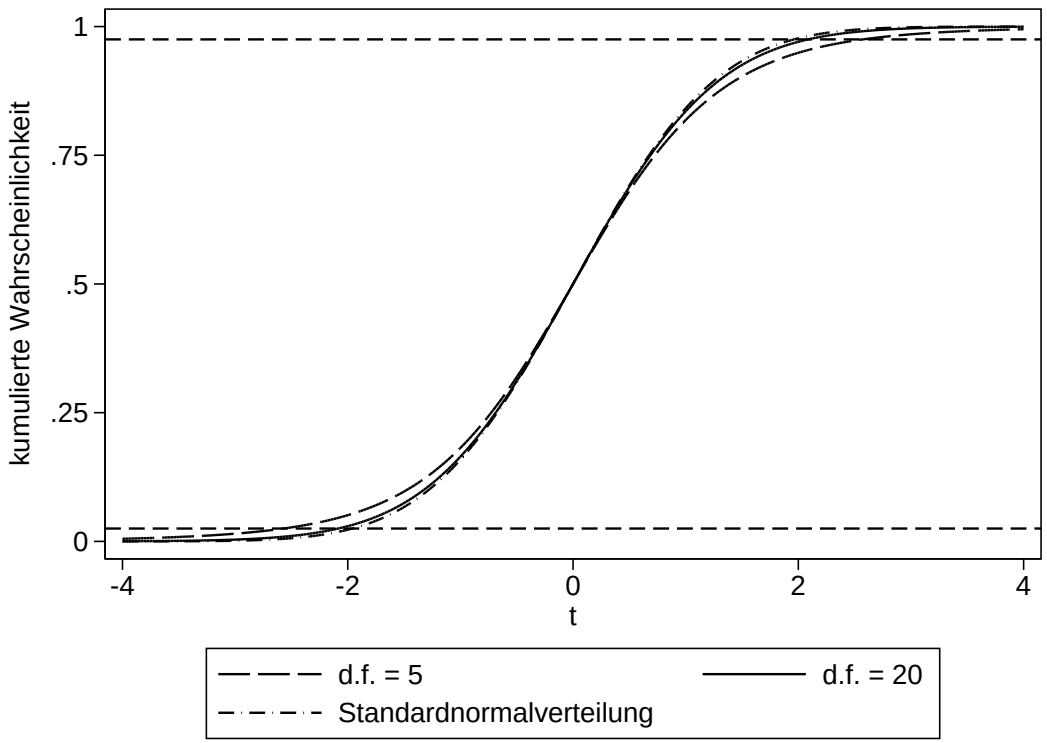
$$T = \frac{Z}{\sqrt{\frac{\chi^2}{df}}}$$

Sie geht bei größerem Stichprobenumfang in die SNV über.

Wichtige Anwendungsfälle: Konfidenzintervall für Mittelwert bei kleinen Stichproben; Vergleich von Mittelwerten; Signifikanztest von Koeffizienten im linearen Regressionsmodell.

Die t-Verteilung illustriert

Verteilungsfunktion einiger t-Verteilungen mit unterschiedlichen Freiheitsgraden (gestrichelt: 0,025- und 0,975-Quantil)



Die F-Verteilung

Die von R. A. Fisher entwickelte (und nach ihm benannte) F-Verteilung entsteht aus zwei von einander unabhängigen χ^2-verteilten Zufallsvariablen:

$$F = \frac{\chi_1^2/df_1}{\chi_2^2/df_2}$$

Im Zähler und im Nenner taucht jeweils ein Freiheitsgrad auf. Es gibt also einen „ersten“ und einen „zweiten“ Freiheitsgrad.

Wichtige Anwendungsfälle: Varianzanalyse (Vergleich von Gruppenmittelwerten, Test des Gesamtmodells im linearen Regressionsmodell).

Was sind Freiheitsgrade?

„Freiheitsgrade“ bezieht sich auf die Zahl der Größen, die frei, d.h. nicht durch andere Größen festgelegt sind.

Beispiele:

- Arithmetisches Mittel und Varianz: Liegt bspw. der Mittelwert fest und stehen n-1 Datenwerte fest, so ist der n.te Datenwert nicht mehr frei $\rightarrow$ n-1 Freiheitsgrade
- Vierfelder-Kreuztabelle: Sind die Randverteilungen und der Wert einer Zelle der Tabelle bekannt, so sind die übrigen drei Werte nicht mehr frei $\rightarrow$ 1 Freiheitsgrad
- Allgemein hat eine Kreuztabelle mit m Zeilen und k Spalten $(m-1)\cdot(k-1)$ Freiheitsgrade